

1/5/3 (Item 3 from file: 351)
DIALOG(R)File 351:Derwent WPI
(c) 2006 The Thomson Corporation. All rts. reserv.

0012868245 - Drawing available
WPI ACC NO: 2002-727258/
XRPX Acc No: N2002-573577

Parallel computer system detects position of node in network based on data transmitted by address conversion entry, when address transmission entry corresponding to transmitted data is selected and registered in cache

Patent Assignee: HITACHI LTD (HITA)

Inventor: KATO S; KAWAMURA T

Patent Family (1 patents, 1 countries)

Patent Application

Number	Kind	Date	Number	Kind	Date	Update
JP 2002259213	A	20020913	JP 200153468	A	20010228	200279 B

Priority Applications (no., kind, date): JP 200153468 A 20010228

Patent Details

Number	Kind	Lan	Pg	Dwg	Filing	Notes
JP 2002259213	A	JA	12	9		

Alerting Abstract JP A

NOVELTY - The data is transmitted between the nodes (102), through the network (101) by converting the address. The position of a node in the network is detected based on the data transmitted by an address conversion entry, when address conversion entry corresponding to the data to be transmitted is selected from address conversion table and registered with the address conversion table cache.

USE - Parallel computer system having several computers connected through network.

ADVANTAGE - The utilization effectiveness of cache is increased and hence speed of converting the address by cache is increased.

DESCRIPTION OF DRAWINGS - The figure shows the block diagram of hardware incorporated in address conversion function and packet division function. (Drawing includes non-English language text).

101 Network
102 Node

Title Terms/Index Terms/Additional Words: PARALLEL; COMPUTER; SYSTEM; DETECT; POSITION; NODE; NETWORK; BASED; DATA; TRANSMIT; ADDRESS; CONVERT; ENTER; TRANSMISSION; CORRESPOND; SELECT; REGISTER; CACHE

Class Codes

International Classification (Main): G06F-012/12
(Additional/Secondary): G06F-012/08, G06F-012/10, G06F-015/16, G06F-015/163

File Segment: EPI;
DWPI Class: T01
Manual Codes (EPI/S-X): T01-H03A

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2002-259213

(P2002-259213A)

(43) 公開日 平成14年9月13日 (2002.9.13)

(51) Int.Cl. ⁷	識別記号	F I	テーマコード(参考)
G 0 6 F 12/12	5 5 5	G 0 6 F 12/12	5 5 5
	5 0 1		5 0 1
12/08	5 0 7	12/08	5 0 7 G
12/10	5 0 1	12/10	5 0 1 F
15/16	6 4 5	15/16	6 4 5

審査請求 未請求 請求項の数 2 O L (全 12 頁) 最終頁に続く

(21) 出願番号 特願2001-53468(P2001-53468)

(22) 出願日 平成13年2月28日 (2001.2.28)

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 加藤 信一

神奈川県秦野市堀山下1番地 株式会社日立製作所エンタープライズサーバ事業部内

(72) 発明者 川村 敏雄

神奈川県秦野市堀山下1番地 株式会社日立製作所エンタープライズサーバ事業部内

(74) 代理人 100075096

弁理士 作田 康夫

Fターム(参考) 5B005 JJ13 KK02 LL11 MM51 NN46

QQ04

5B045 DD12 GG11

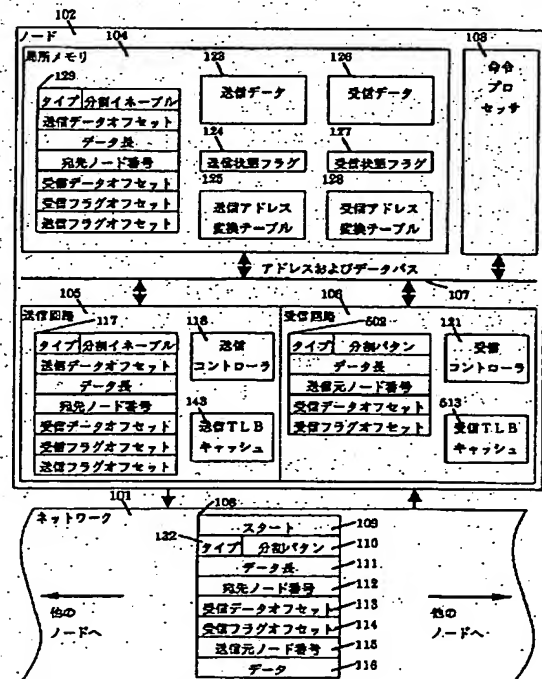
(54) 【発明の名称】 並列コンピュータシステム

(57) 【要約】

【課題】複数のノード間でアドレス変換を行ないながらデータをやり取りするシステムにおいて、複数ページに渡る長いデータを多数転送していると、TLBキャッシュが溢れる可能性が高くなり、TLBキャッシュによるアドレス変換の高速化が期待できなくなる。

【解決手段】TLBキャッシュの各ページエントリに上書き可能であるかないかを示すフィールドを用意し、ページエントリ登録の際に転送するデータの途中のページエントリであればすぐに上書き可能に、またデータの最初か最後であればすぐに上書きできないように優先順位を付ける手段を設ける。

図 1



【特許請求の範囲】

【請求項1】 ネットワークにより接続された複数の処理ノードから構成され、アドレス変換を行ないながらネットワークを介してノード間でデータを転送するコンピュータシステムであって、転送するデータに対応したアドレス変換エントリをアドレス変換テーブルより選択しアドレス変換テーブルキャッシュに登録する際に、アドレス変換エントリの転送するデータに対応する位置を判別可能なことを特徴とする並列コンピュータシステム。

【請求項2】 ネットワークにより接続された複数の処理ノードから構成され、アドレス変換を行ないながらネットワークを介してノード間でデータを転送するコンピュータシステムであって、転送するデータに対応したアドレス変換エントリをアドレス変換テーブルより選択しアドレス変換テーブルキャッシュに登録する際に、アドレス変換エントリの転送するデータに対応する位置によって前記キャッシュ内での優先付けが可能な情報を前記キャッシュのエントリに有することを特徴とする並列コンピュータシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ネットワークによって接続された複数のコンピュータから成るコンピュータシステムに関し、特に多くのコンピュータがローカル及び広域ネットワークを介してパケットを送受信する際にアドレス変換を行なうコンピュータシステムに係る。

【0002】

【従来の技術】複数の計算ノード間のメッセージ通信手段として特開平7-262152号公報の「コンピュータシステム」がある。

【0003】この公知例では送信データアドレス、送信フラグアドレス、受信データアドレス、受信フラグアドレスは局所メモリの実アドレスを示しており、局所メモリ内の連続領域のデータ送受信が可能である。この公知例のようなシステムに対して、局所メモリの実アドレス空間よりも大きなアドレス空間、即ち、仮想アドレス空間でデータの送受信を行なおうとすると、送信回路および受信回路内にアドレス変換機構を設ける必要がある。アドレス変換機構を設けた送信回路や受信回路がデータ送受信時にアドレス変換する際、仮想アドレスと実アドレスとの関係を示しているアドレス変換テーブルにアクセスする必要がある。しかし、仮想アドレスと実アドレスとの関係をきめ細かく行なおうとすると、アドレス変換テーブルのハードウェア物量としては大きくなり、ハードウェアロジックで実現するのは困難である。そこで、局所メモリに設けることで容易に実現可能になるが、アドレス変換テーブルへのアクセス時間が長くなり、アドレス変換のオーバーヘッドが大きくなる。そこで、限られたハードウェアの物量を投入し、アドレス変

換テーブルの一部を格納するTLBキャッシュを備えることでアドレス変換テーブルへのアクセス時間を短くするように工夫している。

【0004】

【発明が解決しようとする課題】一般的なTLBキャッシュの制御方法としてFIFO方式やLRU方式があり、これらの方式を利用してTLBキャッシュを管理する。図9に仮想アドレス空間と実アドレス空間上のデータの例を示す。ここでは転送するデータ701が仮想アドレス空間702と実アドレス空間703との関係を示すアドレス変換単位をページ704と呼ぶ。ページ番号と実アドレス空間の先頭アドレスとの関係を記述したエントリを複数備えたアドレス変換テーブルにより、仮想アドレス空間702のページ704はこのように自由に実アドレス空間703に割り当てることが可能である。この例ではデータ701はページ番号1から4までの広範囲に渡っており、実アドレスに変換する際に、TLBキャッシュに対して4ページ分ものページエントリの登録が必要になる。TLBキャッシュには限られた数のページエントリしか格納できないので、ページ704を複数に渡るような長いデータ701を多数転送していると、直ぐに一杯になってしまい溢れる。ページ番号1や4などのデータ転送にはまだ使用されていない領域があって直ぐに再利用される可能性の高いページエントリと、ページ番号2や3などのデータ転送に全て使用されたページエントリは分別なく扱われるため、再利用される前に別のデータ転送によって溢れる可能性が高くなる。TLBキャッシュのページエントリを再利用する前に溢れて消えてしまうと、再び局所メモリにアクセスしなければならず、TLBキャッシュによるアドレス変換の高速化が期待できなくなる。

【0005】

【課題を解決するための手段】上記目的を達成するために、TLBキャッシュ内の全てのエントリに、上書きを優先してもよいかどうかを示すオーバーライトフィールドを設ける。次に、局所メモリにあるアドレス変換テーブルのエントリをTLBキャッシュに登録するときに、オーバーライトフィールドが有効であるエントリがあればそのエントリに上書きする手段を設ける。さらに、ページエントリをTLBキャッシュに登録後、そのページエントリがデータの最初のエントリであるか途中のエントリであるか、あるいは最後のエントリであるかを判別する手段を設けて、データの途中のページエントリであるときはオーバーライトフィールドを有効にし、データの最初か最後のエントリであればオーバーライトフィールドを無効にする手段を設ける。このようにオーバーライトフィールドを設けることで、直ちに再利用される可能性の低いページエントリから先に後続のデータのページエントリに上書きされ、再利用される可能性の高いページエントリを優先してTLBキャッシュ内に残すこと

ができるので、複数ページにまたがる長いデータを多数転送するシステムにおいて有効である。

【0006】

【発明の実施の形態】本実施例を図1に示す。これはネットワーク101で接続した複数のノードからなる。便宜上、ノード102の内の1個のみを図1に示す。図示のノード102は送信ノードと受信ノードの両方を示すために用いられているが、実際にはこれらのノードは別のものである。各ノード102は命令プロセッサ103、適当な容量を備えた局所メモリ104、送信回路105および受信回路106を含む。アドレスおよびデータバス107は送信回路105にコマンドを出し、命令プロセッサ103と局所メモリ104の間にデータを転送し、そして局所メモリ104と送信回路105および受信回路106との間でデータを転送するために用いられる。これらの4つの接続はアドレスおよびデータバス107で実現されているが、それぞれ独立のアドレス/データ線でもよい。

【0007】ネットワーク101はノード102内の送信回路105からパケット108dを受け取り、各パケット108に含まれる宛先ノード番号112の内容に従って他のノード102の受信回路106にそれらを送る。本例ではネットワーク101はパケット108を送信順に受信回路106に転送する。

【0008】図2は送信回路105のブロック図である。送信シーケンサ130は送信回路105全体の制御を行なう。送信コマンドは送信制御ワードブロック129（以下、TCWBと呼ぶ）の形をとり、これが命令プロセッサ103により局所メモリ104に記憶される。

【0009】TCWBアドレスレジスタ131は命令プロセッサ103により書き込み可能なレジスタである。TCWBバッファ117はTCWB129のコピーを一時的に保持するために用いられるレジスタファイルである。これはTCWBと同じフォーマットを有し、そしてパケットタイプ132、パケット分割イネーブル133、送信データオフセット134、データ長111、宛先ノード番号112、受信データオフセット113、受信フラグオフセット114、送信フラグオフセット135を含む。

【0010】送信FIFOデータバッファ136はファーストイン・ファーストアウト（FIFO）データバッファであって、局所メモリ104とネットワーク101間のデータの緩衝に用いられる。送信直接メモリアクセスコントローラ137（以下、送信DMACと呼ぶ）は送信状態フラグ124を局所メモリ104に書き込むことができ、あるいは局所メモリ104からのブロック読み出しを行なうことのできる回路である。このブロック読み出し動作のために、送信DMAC137はそのブロックの長さおよびそのブロックのスタートアドレスに初期化される長さカウンタとアドレスカウンタを含む。

【0011】送信DMAC137へのコマンドは送信アドレス変換回路138と送信シーケンサ130から得られる。ブロック読み出し動作については、アドレスおよびデータバス107に出力されるアドレスカウンタの現在の値を用いて局所メモリ読み出し要求が発生され、アドレスカウンタが増分され長さが0となるまで長さカウンタから減分される。

【0012】局所メモリ104から読み出されたデータは送信シーケンサ130の制御によりTCWBバッファ117または送信FIFOデータバッファ136に記憶される。送信ノードレ番号レジスタ139は各ノード102に固有の、そのノードを識別する番号を含む。

【0013】ページサイズレジスタ140は仮想アドレス空間を複数のページで分割したときの1ページあたりのページサイズを示している。TLBアドレスレジスタ141は局所メモリ104にある送信アドレス変換テーブル125の先頭アドレスを示している。

【0014】パケット組み立て部142はTCWBバッファ117を送信ノード番号レジスタ139から情報を抽出してパケットヘッダを構成し、送信FIFOデータバッファ136からデータを取り出してパケット108を形成し、そして完全なパケット108をネットワーク101に送り出す回路である。

【0015】ノード102について1つ以上のパケット108を1つ以上の他のノード102に送るために、送信ノード102内の命令プロセッサ103は、まず転送されるべき各パケット108についての情報として、局所メモリ104にTCWB129を作る。

【0016】この命令プロセッサ104はTCWB129内に含まれる各フィールドを初期化する。タイプ132は初期化されたパケットタイプを示す。本例では全てのパケットタイプは同じであるからタイプ132は固定値である。

【0017】分割イネーブル133は1つのパケットで送信するデータ長を制限するかどうかを指定している。送信データオフセット134は仮想アドレス空間上の送信データの先頭アドレスを示しており、局所メモリ104から送信データを読み出すときには実アドレスに変換する必要があるが、これについては後述する。データ長111は転送されるべきデータの長さを示す。宛先ノード番号112はそのパケットについての宛先ノードを識別する。受信データオフセット113は仮想アドレス空間上の受信データの先頭アドレスを示しており、局所メモリ104に受信データを書き込むときには実アドレスに変換する必要があるが、これについては後述する。受信フラグオフセット114は宛先ノードの局所メモリ104に、受信状態フラグ127をポイントすべきところを示すが、ポイントするには実アドレスに変換する必要がある、これについては後述する。

【0018】パケット送信プロセスをスタートするため

に、命令プロセッサ104はアドレスおよびデータバス107を用いてTCWB129の先頭アドレスをTCWBアドレスレジスタ131に書き込む。送信シーケンサxxxはこのレジスタへの書き込みにより動作を開始する。

【0019】まず、送信シーケンサ130はTCWBアドレスレジスタ131に書き込まれたアドレスを用いて送信DMAC137にTCWB129の読み出しを命令する。TCWB129はアドレスおよびデータバス107を通して局所メモリ104から読み出され、そしてTCWBバッファ117に記憶される。次に、送信シーケンサ130は送信アドレス変換回路138に対して送信データの読み出しを行なうよう命令する際、オフセット情報と送信データ長情報、分割パタン情報の3つを渡す必要がある。しかし、TCWBバッファ129内の分割イネーブル133が有効であるか無効であるかによって値が異なる。

【0020】分割イネーブル133が無効である場合、本例では0'であるとき、送信アドレス変換回路138に渡すオフセット情報と送信データ長情報、分割パタン情報はそれぞれ次の通りとなる。

(オフセット情報) = (TCWBバッファ117内の送信データオフセット134)

(送信データ長情報) = (データ長111)

(分割パタン情報) = 00'

分割イネーブル133が有効である場合、本例では1'であるとき、TCWBバッファ117内のデータ長111が分割パケット長(本例では固定値)より大きい小さいかによって、送信シーケンサ130の処理が2つに分けられる。

【0021】データ長111が分割パケット長と同じか小さい場合、分割イネーブル133が0'であるときと同様、TCWBバッファ117内の送信データオフセット134とデータ長111、分割パタン情報00'を送信アドレス変換回路138に渡す。

【0022】データ長111が分割パケット長より大きい場合、パケットのデータ長が分割パケット長より大きくならないように複数のパケットに分けて送信する。そのため、送信シーケンサ130は送信アドレス変換回路138に対して、複数に分割して送信データ読み出し命令を行なう。

【0023】分割の最初の送信データの読み出し命令では、オフセット情報と送信データ長情報、分割パタン情報がそれぞれ次の通りとなる。

(オフセット情報) = (TCWBバッファ117内の送信データオフセット134)

(送信データ長情報) = 分割パケット長

(分割パタン情報) = 10'

分割途中の送信データの読み出し命令では、オフセット情報と送信データ長情報、分割パタン情報がそれぞれ次

の通りとなる。

(オフセット情報) = (TCWBバッファ117内の送信データオフセット134) + ((分割パケット長) × ((分割開始からの順番) - 1))

(送信データ長情報) = 分割パケット長

(分割パタン情報) = 11'

分割の最後の送信データの読み出し命令では、オフセット情報と送信データ長情報、分割パタン情報がそれぞれ次の通りとなる。

(オフセット情報) = (TCWBバッファ117内の送信データオフセット134) + ((分割パケット長) × ((分割開始からの順番) - 1))

(送信データ長情報) = (送信データ長を分割パケット長で割った余り)

(分割パタン情報) = 01'

図5に仮想アドレス空間と実アドレス空間の関係を示す。仮想アドレス空間149は複数のページ150に別れていて、ページ単位で実アドレス空間151上の連続領域に割り当てられている。ページの大きさはページサイズレジスタ140により決まる。なお、送信データ123の大きさはページサイズに無関係のため、複数のページ150に渡る場合があるが、本例では、送信データ123が複数のページ150に渡らない場合について記述している。該当するページ番号が判別したら、局所メモリ104にある送信アドレス変換テーブル125を参照して実アドレスに変換する。送信アドレス変換テーブル125は、局所メモリ104上の連続アドレス領域にページ番号0、1、2...n番まで順番に並べており、送信アドレス変換テーブル125の先頭アドレスはTLBアドレスレジスタ141に示されている。仮想アドレス空間上の送信データの先頭アドレスはTCWB129内の送信データオフセット134で示されており、送信データ123がどのページに存在しているかが判別可能である。送信アドレス変換テーブルから、該当するページ150の先頭実アドレスを求める。さらに、送信データオフセット134をページサイズで割った余りがページ内の送信データオフセットであり、このオフセットを先に求めた該当するページ150の先頭実アドレスに加算した値が送信データの先頭実アドレスとなる。

【0024】送信アドレス変換回路138は、局所メモリ104にある送信データの先頭実アドレスと送信データ長を求めて送信DMAC137にデータを読み出す指示と、仮想アドレスから実アドレスへの変換の効率向上のために使用する送信TLBキャッシュ143の管理を行なう。詳細は図3、図4に示す手順の通りに行なう。

【0025】第1に、送信アドレス変換回路138は、仮想アドレス空間上のページ番号を求めるため、オフセット情報をページサイズレジスタの値で割って商を計算する。

【0026】第2に、送信アドレス変換回路138は送

信TLBキャッシュ143に、先に求めたページ番号が存在するかどうかを確認する。送信TLBキャッシュ143は図2に示すように、バリッド145、オーバーライト146、ページ番号147、ページ内先頭実アドレス148の4つの要素を1つのエントリ144として複数エントリ144が存在する。送信TLBキャッシュ143の全エントリ144を見て、バリッド145が有効でかつページ番号が同一のエントリ144があるかどうかを探す。エントリ144が存在した場合、第10項の処理へ移る。

【0027】第3に、同一のエントリ144が無い場合、局所メモリ104の送信アドレス変換テーブル125内にあるエントリ144の先頭アドレスを求める必要がある。先に求めたページ番号にTLBエントリサイズ、本例では固定値を掛け算し、TLBアドレスレジスタ141の値を加算する。

【0028】第4に、エントリ144の先頭アドレスを求めたら、送信DMAC137に対してエントリを局所メモリ104から読み出すように命令する。

【0029】第5に、エントリが局所メモリ104から読み出されるのを待つ。

【0030】第6に、送信TLBキャッシュ143にあるエントリ144に上書き可能なエントリ144があるかどうかを確認する。具体的には、送信TLBキャッシュ143にあるエントリ144の中で、オーバーライト146が1'のエントリが存在するかどうかを確認する。

【0031】第7に、送信TLBキャッシュ143にあるエントリ144の中で、バリッド145が1'でかつオーバーライト146が1'であるエントリ144が存在する場合、そのエントリ144に対して、局所メモリ104より読み出されたエントリを上書きする。具体的には、先の計算で求めたページ番号と局所メモリ104より読み出されたエントリにある該当ページの先頭実アドレスを書き込む。バリッド145が1'でかつオーバーライト146が1'であるエントリ144が存在しない場合、FIFO方式やLRU方式などの規則によってTLBエントリを1つ選択する。選択したTLBエントリのバリッド145が1'の場合、読み出されたエントリを上書きし、バリッド145が0'であれば新たに読み出されたエントリを登録する。エントリの登録とは、前述の上書き時の処理に加えてさらに、バリッド145を1'に設定することを指す。バリッド145とはエントリ144そのものが有効であるか無効であるかを示しており、一般的にFIFO方式やLRU方式などで管理されたエントリ144の情報である。

【0032】第8に、上書きまたは登録したエントリ144に対して、上書き設定をどうするかを判定する。上書き設定とは、そのエントリ144のオーバーライト146を設定することを指す。次の4つの判定条件があ

る。

条件1 分割パタン情報が00'で送信データの最初のページのエントリである。

条件2 分割パタン情報が00'で送信データの最後のページのエントリである。

条件3 分割パタン情報が10'で送信データの最初のページのエントリである。

条件4 分割パタン情報が01'で送信データの最後のページのエントリである。

10 【0033】第9に、上記の4つの条件のうち、いずれも成立しない場合、オーバーライト146を1'設定する。いずれかが成立した場合、オーバーライト146を0'に設定する。上記条件1~2により、分割されないパケットの送信データの最初と最後のページのエントリのオーバーライト146を0'に設定することにより、前述の第7項での上書きが行われなくなるが、途中のページのエントリのオーバーライト146を1'に設定することにより、第7項での上書きが行われることになる。また、上記条件3~4により、分割された先頭パケットの送信データの最初のページのエントリと、分割された末尾パケットの送信データの最初のページのエントリのオーバーライト146を0'に設定することにより、第7項での上書きが行われなくなるが、分割される前の連続した送信データとして着目すると途中のページのエントリのオーバーライト146を1'に設定することにより、第7項での上書きが行われることになる。

【0034】第10に、第4項で求めたページ番号のTLBエントリ内にある該当ページの先頭実アドレスを使って、送信データの先頭実アドレスを求める。送信データの先頭アドレス情報をページサイズで割り、その余りを該当ページの先頭実アドレスに加算した値が送信データの先頭実アドレスである。

【0035】このようにして、送信アドレス変換回路138は、送信データの先頭実アドレスを求め、送信データ長情報とともに送信DMAC137に渡して、データの読み出しを命令する。

【0036】送信シーケンサは、局所メモリ104より読み出された送信データをFIFOデータバッファ141に記憶し、そして次に分割パタン情報をパケット組み立て部142に渡して、パケット108の送信開始を命令する。

【0037】パケット組み立て部142は、まずパケット108のスタートをマークする固定スタートコード109を送る。また、パケット組み立て部142は、TCWBバッファ117と送信ノード番号レジスタ139、送信シーケンサ130から渡される分割パタン情報の3つから、タイプ132、分割パタン110、宛先ノード番号112、受信データオフセット113、受信フラグオフセット114、送信元ノード番号115を取り出してそれを送信する。

【0038】次に、パケット組み立て部142は送信FIFOデータバッファ136から送信データを取り出し、それを送ってパケットデータ116を形成する。なお、送信FIFOデータバッファ136はこれと並行して送信DMAC137により局所メモリ104から読み出された送信データを記憶している。

【0039】送信データを全て送信し終わると、パケット組み立て部142は送信シーケンサ130にパケット送信処理完了を通知する。送信シーケンサ130は、送信アドレス変換回路138にTCWBバッファ117に記憶された送信フラグオフセット135を渡して、そのパケットが送信されたことを示す値を送信状態フラグ124に書き込むことを命令する。本例では送信状態フラグ124は、パケット108が送信されたか、まだ送信されていないかを意味する2つの可能な値の1つを有することができる。

【0040】本例では、送信状態フラグ124は仮想アドレス空間上のページサイズよりも小さい大きさで、さらに、2つのページをまたぐことが無いように、あらかじめ送信フラグオフセット135の値を制限されているので、送信データの読み出し命令のときと同様に、送信状態フラグ124が複数のページに渡らない場合について記述する。

【0041】第1に、送信アドレス変換回路138は、仮想空間上のページ番号を求めるため、送信フラグオフセットをページサイズレジスタの値で割って商を計算する。

【0042】第2に、送信アドレス変換回路138は送信TLBキャッシュ143に、先に求めたページ番号が存在するかどうかを確認する。具体的には、送信TLBキャッシュ143の全エントリ144を見て、バリッド145が有効でページ番号が同一のエントリ144があるかどうかを探す。エントリ144が存在した場合、第9項の処理へ移る。

【0043】第3に、同一のエントリが無い場合、局所メモリ104にあるエントリアドレスを求めるため、先に求めたページ番号にエントリサイズ、この実施例では固定値を掛け算し、TLBアドレスレジスタ141の値を加算する。

【0044】第4に、エントリアドレスを求めたら、送信DMAC137に対してエントリアドレスの指すエントリを局所メモリ104から読み出すように命令する。

【0045】第5に、エントリが局所メモリ104から読み出されるのを待つ。

【0046】第6に、送信TLBキャッシュ143にあるエントリ144に上書き可能なエントリ144があるかどうかを確認する。具体的には、送信TLBキャッシュ143にあるエントリ144の中で、オーバーライト146が1'のエントリが存在するかどうかを確認する。

【0047】第7に、送信TLBキャッシュ143にあるエントリの中で、バリッド145が1'でかつオーバーライト146が1'であるエントリが存在する場合、そのエントリ144に対して、局所メモリ104より読み出されたエントリを上書きする。具体的には、先の計算で求めたページ番号と読み出されたエントリにある該当ページの先頭実アドレスを書き込む。バリッド145が1'でかつオーバーライト146が1'であるエントリ144が存在しない場合、FIFO方式やLRU方式などの規則によってエントリ144を1つ選択する。選択したエントリ144のバリッド145が1'の場合、読み出されたエントリを上書きし、バリッド145が0'であれば新たに読み出されたエントリ144を登録する。エントリ144の登録とは、前述の上書き時の処理に加えてさらに、バリッド145を1'に設定することを指す。

【0048】第8に、上書きまたは登録したエントリ144に対して、上書き設定する。上書き設定とは、そのエントリのオーバーライト146を設定することを指す。送信状態フラグ124は先に述べた通りページサイズよりも小さく、再利用される可能性があるためオーバーライト146は0'に設定する。

【0049】第9に、第4項で求めたページ番号のエントリ144内にある該当ページの先頭実アドレス148を使って、送信状態フラグ124の先頭実アドレスを求める。送信フラグオフセット135をページサイズで割り、その余りを該当ページの先頭実アドレス148に加算した値が送信状態フラグ124の先頭実アドレスである。

【0050】このようにして、送信アドレス変換回路138は、送信状態フラグ124の先頭実アドレスを求め、送信DMAC137に渡して、送信状態フラグ124の書き込みを命令する。

【0051】ネットワーク101は送信回路105からのパケット108を受け、そして宛先ノード番号112に従い、宛先ノードの受信回路106にそのパケット108を転送する。本例のネットワーク101は、複数のパケット108が1つのノードから他のノードに送られるときは、それらのパケット108が送信順に入るように構成されている。

【0052】図6は受信回路106のブロック図である。受信シーケンサ501はパケット受信状態フラグ127の書き込み処理を含む受信回路106全体の制御を行なう。パケットヘッダバッファ502はパケットヘッダのコピーを一時的に記憶するために用いられる。このバッファ502のフォーマットはパケットヘッダのそれとほとんど同じであり、それはパケットタイプ503、分割ボタン504、データ長505、受信データオフセット506、受信フラグオフセット507、送信元ノード番号508を含む。

【0053】パケット受信部509はパケット101が到着すると、到着したことを受信シーケンサ501に伝えたとともに、パケットヘッダをパケットヘッダバッファ502へ、受信データを受信FIFOデータバッファ510に格納する。受信FIFOデータバッファ510はネットワーク101からのデータを受けて、そのデータが局所メモリ104に書き込まれる前に緩衝するバッファである。

【0054】受信直接メモリアクセスコントローラ512（以下、受信DMACと呼ぶ）は受信状態フラグ127を局所メモリ104に書き込むことができ、あるいは局所メモリ104にブロック書き込みを行なうことのできる回路である。このブロック書き込み動作のために、受信DMAC512はそのブロックの長さおよびそのブロックのスタートアドレスに初期化される長さカウンタとアドレスカウンタを含む。

【0055】受信DMAC512へのコマンドは受信アドレス変換回路511と受信シーケンサ501から得られる。ブロック書き込み動作については、アドレスおよびデータバス107に出力されるアドレスカウンタの現在の値を用いて局所メモリ書き込み要求が発生され、アドレスカウンタが増分され長さが0となるまで長さカウンタから減分される。

【0056】以下、受信回路106の動作を説明する。受信回路106内のパケット受信部509はパケット108のスタートマーク109を受け取ると、受信シーケンサ501にパケットの到着を知らせる。受信したパケットの残りの部分からのマークは分離して捨てる。パケットヘッダのうち、ヘッダバッファ502に該当するフィールドのみをヘッダバッファ502に格納し、パケットヘッダのデータ長111を用いて受信FIFOデータバッファ510にデータを格納する。

【0057】受信シーケンサ501は受信アドレス変換回路511に対して受信データの書き込みを行なうよう命令する際、オフセット情報と受信データ長情報、分割パタン情報の3つを渡す。3つの情報はそれぞれ次の通りとなる。

（オフセット情報）＝（パケットヘッダバッファ502内の受信データオフセット506）

（受信データ長情報）＝（受信データ長505）

（分割パタン情報）＝（パケットヘッダバッファ502内の分割パタン504）

受信アドレス変換回路511は、局所メモリ104にある受信データの先頭実アドレスと受信データ長を求めて受信DMAC512にデータを書き込む指示と、仮想アドレスから実アドレスへの変換の効率向上のために使用する受信TLBキャッシュ513の管理を行なう。詳細は図7、図8に示す手順の通りに行なう。

【0058】第1に、受信アドレス変換回路511は、仮想アドレス空間上のページ番号を求めるため、オフセ

ット情報をページサイズレジスタの値で割って商を計算する。

【0059】第2に、受信アドレス変換回路511は受信TLBキャッシュ513に、先に求めたページ番号が存在するかどうかを確認する。受信TLBキャッシュ513は図6に示すように、バリッド517、オーバーライト518、ページ番号519、ページ内先頭実アドレス520の4つの要素を1つのエントリ516として複数エントリ516が存在する。受信TLBキャッシュ513の全エントリ516を見て、バリッド517が有効でかつページ番号が同一のエントリ516があるかどうかを探す。エントリ516が存在した場合、第10項の処理へ移る。

【0060】第3に、同一のエントリ516が無い場合、局所メモリ104の受信アドレス変換テーブル128内にあるエントリの先頭アドレスを求める必要がある。先に求めたページ番号にTLBエントリサイズ、本例では固定値を掛け算し、TLBアドレスレジスタ515の値を加算する。

【0061】第4に、エントリの先頭アドレスを求めたら、受信DMAC512に対してエントリを局所メモリ104から読み出すように命令する。

【0062】第5に、エントリが局所メモリ104から読み出されるのを待つ。

【0063】第6に、受信TLBキャッシュ513にあるエントリ516に上書き可能なエントリ516があるかどうかを確認する。具体的には、受信TLBキャッシュ513にあるエントリ516の中で、オーバーライト518が1'のエントリが存在するかどうかを確認する。

【0064】第7に、受信TLBキャッシュ513にあるエントリ516の中で、バリッド517が1'でかつオーバーライト518が1'であるエントリ516が存在する場合、そのエントリ516に対して、局所メモリ104より読み出されたエントリを上書きする。具体的には、先の計算で求めたページ番号と局所メモリ104より読み出されたエントリにある該当ページの先頭実アドレスを書き込む。バリッド517が1'でかつオーバーライト518が1'であるエントリ516が存在しない場合、FIFO方式やLRU方式などの規則によってTLBエントリを1つ選択する。選択したTLBエントリのバリッド517が1'の場合、読み出されたエントリを上書きし、バリッド517が0'であれば新たに読み出されたエントリを登録する。エントリの登録とは、前述の上書き時の処理に加えてさらに、バリッド517を1'に設定することを指す。バリッド517とはエントリ516そのものが有効であるか無効であることを示しており、一般的にFIFO方式やLRU方式などで管理されたエントリ516の情報である。

【0065】第8に、上書きまたは登録したエントリ5

16に対して、上書き設定をどうするかを判定する。上書き設定とは、そのエントリ516のオーバーライト518を設定することを指す。次の4つの判定条件がある。

条件1 分割パタン情報が00'で受信データの最初のページのエントリである。

条件2 分割パタン情報が00'で受信データの最後のページのエントリである。

条件3 分割パタン情報が10'で受信データの最初のページのエントリである。

条件4 分割パタン情報が01'で受信データの最後のページのエントリである。

【0066】第9に、上記の4つの条件のうち、いずれも成立しない場合、オーバーライト518を1'設定する。いずれかが成立した場合、オーバーライト518を0'に設定する。上記条件1~2により、分割されないパケットの受信データの最初と最後のページのエントリのオーバーライト518を0'に設定することにより、前述の第7項での上書きが行われなくなるが、途中のページのエントリのオーバーライト518を1'に設定することにより、第7項での上書きが行われることになる。また、上記条件3~4により、分割された先頭パケットの受信データの最初のページのエントリと、分割された末尾パケットの受信データの最初のページのエントリのオーバーライト518を0'に設定することにより、第7項での上書きが行われなくなるが、分割される前の連続した受信データとして着目すると途中のページのエントリのオーバーライト518を1'に設定することにより、第7項での上書きが行われることになる。

【0067】第10に、第4項で求めたページ番号のTLBエントリ内にある該当ページの先頭実アドレスを使って、受信データの先頭実アドレスを求める。受信データの先頭アドレス情報をページサイズで割り、その余りを該当ページの先頭実アドレスに加算した値が受信データの先頭実アドレスである。

【0068】このようにして、受信アドレス変換回路511は、受信データの先頭実アドレスを求め、受信データ長情報とともに受信DMAC1512に渡して、データの書き込みを命令する。

【0069】受信DMAC512内の長さカウンタが0になると、受信データの全てが局所メモリ104に書き込まれたことになる。受信シーケンサ501は受信アドレス変換回路511にパケットヘッダバッファ502に記憶された受信フラグオフセット507を渡して、そのパケットが受信されたことを示す値を受信状態フラグ127の書き込みを命令する。本例では受信状態フラグ127は、パケット108が受信されたか、まだ受信されていないかを意味する2つの可能な値の1つを有することができる。本例では、受信状態フラグ127は仮想アドレス空間上のページサイズよりも小さい大きさで、さ

らに、2つのページをまたぐことが無いように、あらかじめ受信フラグオフセット507の値を制限されているので、受信データの書き込み命令のときと同様に、受信状態フラグ127が複数のページに渡らない場合について記述する。

【0070】第1に、送信アドレス変換回路138は、仮想空間上のページ番号を求めるため、送信フラグオフセットをページサイズレジスタの値で割って商を計算する。

10 【0071】第2に、受信アドレス変換回路511は受信TLBキャッシュ513に、先に求めたページ番号が存在するかどうかを確認する。具体的には、受信TLBキャッシュ513の全エントリ516を見て、バリッド517が有効でページ番号が同一のエントリ516があるかどうかを探す。エントリ516が存在した場合、第9項の処理へ移る。

20 【0072】第3に、同一のエントリが無い場合、局所メモリ104にあるエントリアドレスを求めるため、先に求めたページ番号にエントリサイズ、この実施例では固定値を掛け算し、TLBアドレスレジスタ515の値を加算する。

【0073】第4に、エントリアドレスを求めたら、受信DMAC512に対してエントリアドレスの指すエントリを局所メモリ104から読み出すように命令する。

【0074】第5に、エントリが局所メモリ104から読み出されるのを待つ。

30 【0075】第6に、受信TLBキャッシュ513にあるエントリ516に上書き可能なエントリ516があるかどうかを確認する。具体的には、受信TLBキャッシュ513にあるエントリ516の中で、オーバーライト517が1'のエントリが存在するかどうかを確認する。

40 【0076】第7に、受信TLBキャッシュ513にあるエントリの中で、バリッド517が1'でかつオーバーライト518が1'であるエントリが存在する場合、そのエントリ516に対して、局所メモリ104より読み出されたエントリを上書きする。具体的には、先の計算で求めたページ番号と読み出されたエントリにある該当ページの先頭実アドレスを書き込む。バリッド517が1'でかつオーバーライト518が1'であるエントリ516が存在しない場合、FIFO方式やLRU方式などの規則によってエントリ516を1つ選択する。選択したエントリ516のバリッド517が1'の場合、読み出されたエントリを上書きし、バリッド517が0'であれば新たに読み出されたエントリ516を登録する。エントリ516の登録とは、前述の上書き時の処理に加えてさらに、バリッド517を1'に設定することを指す。

50 【0077】第8に、上書きまたは登録したエントリ516に対して、上書き設定する。上書き設定とは、その

エントリのオーバーライト518を設定することを指す。受信状態フラグ127は先に述べた通りページサイズよりも小さく、再利用される可能性があるためオーバーライト518は0'に設定する。

【0078】第9に、第4項で求めたページ番号のエントリ516内にある該当ページの先頭実アドレス520を使って、受信状態フラグ127の先頭実アドレスを求める。受信フラグオフセット506をページサイズで割り、その余りを該当ページの先頭実アドレス520に加算した値が受信状態フラグ127の先頭実アドレスである。

【0079】このようにして、受信アドレス変換回路511は、受信状態フラグ127の先頭実アドレスを求め、受信DMAC512に渡して、受信状態フラグ127の書き込みを命令する。

【0080】宛先ノードの命令プロセッサ103は任意の時点でパケット108の受信状態フラグ127を読み取るところにより、そのパケット108の受信を確認できる。

【0081】本発明の実施例は、連続したデータを複数のノード間で転送する際に、利用効率の低いアドレス変換テーブルエントリをTLBキャッシュに常駐させないようにして、TLBキャッシュの利用効率を上げ、アドレス変換を高速化するための手段を与えるものである。

【0082】

【発明の効果】本発明の並列コンピュータシステムではTLBキャッシュの各ページエントリに上書き可能であるかないかを示すフィールドを設けて、ページエントリ登録の際に転送するデータの途中のページエントリであればすぐに上書き可能に、またデータの最初か最後であればすぐに上書きできないように優先順位を付けることが可能になる。これにより、データの最初または最後のページに相当する、データ転送にはまだ使用されていない領域があって再利用される可能性の高いページエントリと、データ転送の際に全ての領域を使用されたページエントリを区別して扱えるため、再利用される前に別のデータ転送によって溢れる可能性が低くなりTLBキャッシュの利用効率が上がってアドレス変換の高速化が期待できる。

【図面の簡単な説明】

【図1】アドレス変換機能およびパケット分割機能を組み込んだハードウェアのブロック図。

【図2】本発明を組み込んだ送信回路のブロック図。

【図3】本発明を組み込んだ送信アドレス変換回路における送信データのアドレス変換処理手順を示す図。

【図4】図3のaに続く送信データのアドレス変換処理手順を示す図。

【図5】送信データのアドレスを変換するときの仮想ア

ドレス空間と実アドレス空間およびアドレス変換テーブルとの関係を示した図。

【図6】本発明を組み込んだ受信回路のブロック図。

【図7】本発明を組み込んだ受信アドレス変換回路における受信データのアドレス変換処理手順を示す図。

【図8】図7のaに続く受信データのアドレス変換処理手順を示す図。

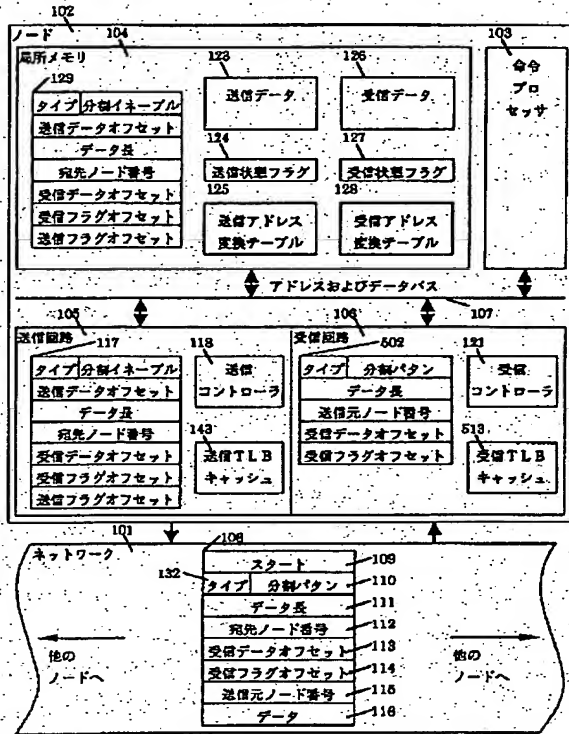
【図9】従来の転送するデータのアドレスを変換するときの仮想アドレス空間と実アドレス空間の関係を示した図。

【符号の説明】

101…ネットワーク、102…ノード、103…命令プロセッサ、104…局所メモリ、105…送信回路、106…受信回路、107…アドレスおよびデータバス、108…パケット、109…スタートコード、110…分割パタン、111…データ長、112…宛先ノード番号、113…受信データオフセット、114…受信フラグオフセット、115…送信先ノード番号、116…データ、117…TCWBバッファ、118…送信コントローラ、121…受信コントローラ、123…送信データ、124…送信状態フラグ、125…送信アドレス変換テーブル、126…受信データ、127…受信状態フラグ、128…受信アドレス変換テーブル、129…転送制御ワードブロック(TCWB)、131…TCWBアドレスレジスタ、132…タイプ、133…分割インネブル、134…送信データオフセット、135…送信フラグオフセット、136…送信FIFOデータバッファ、137…送信DMAC、138…送信アドレス変換回路、139…送信ノード番号レジスタ、140…ページサイズレジスタ、141…TLBアドレスレジスタ、142…パケット組み立て部、143…送信TLBキャッシュ、144…ページエントリ、145…バリッド、146…オーバーヘッド、147…ページ番号、148…ページ内先頭実アドレス、149…仮想アドレス空間、150…ページ、151…実アドレス空間、501…受信シーケンサ、502…パケットヘッダバッファ、503…タイプ、504…分割パタン、505…データ長、506…受信データオフセット、507…受信フラグオフセット、508…送信元ノード番号、509…パケット受信部、510…受信FIFOデータバッファ、511…受信アドレス変換回路、512…受信DMAC、513…受信TLBキャッシュ、514…ページサイズレジスタ、515…TLBアドレスレジスタ、516…ページエントリ、517…バリッド、518…オーバーライト、519…ページ番号、520…ページ内先頭実アドレス、701…データ、702…仮想アドレス空間、703…実アドレス空間、704…ページ。

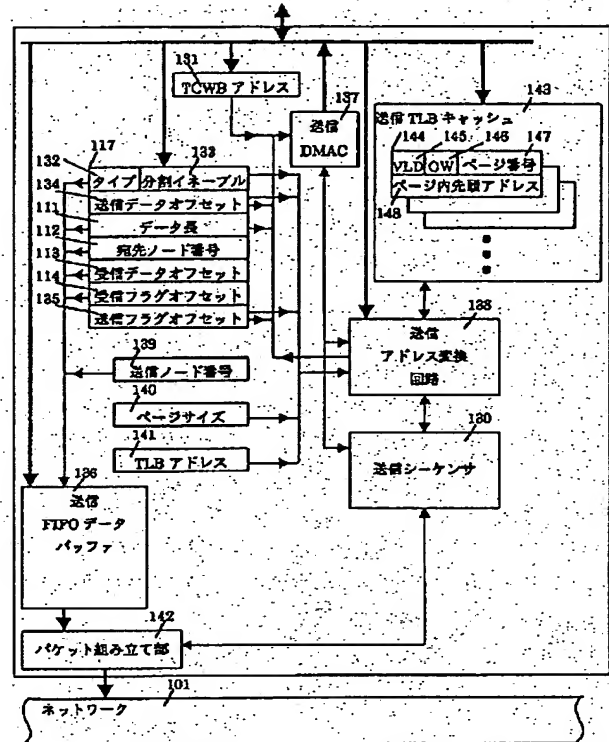
【図1】

図 1



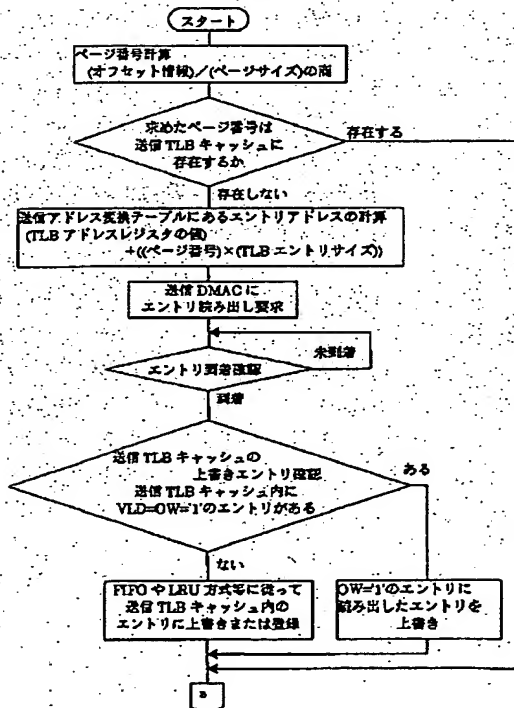
【図2】

図 2



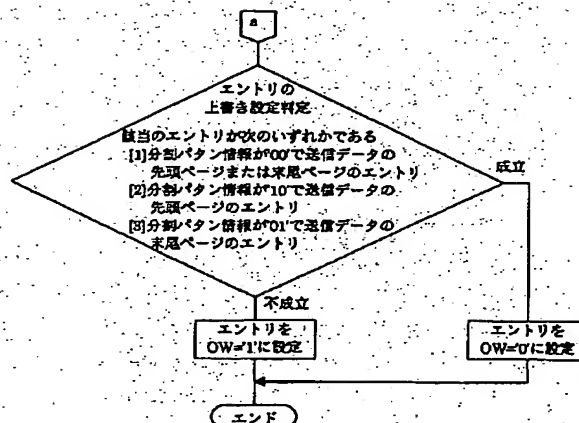
【図3】

図 3



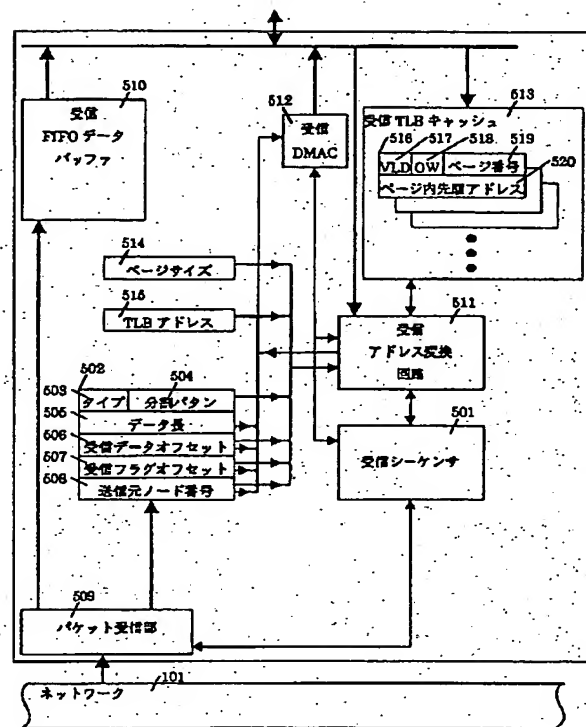
【図4】

図 4



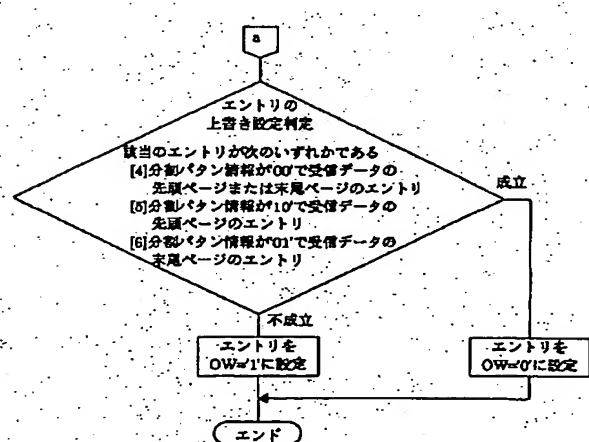
【図 6】

6



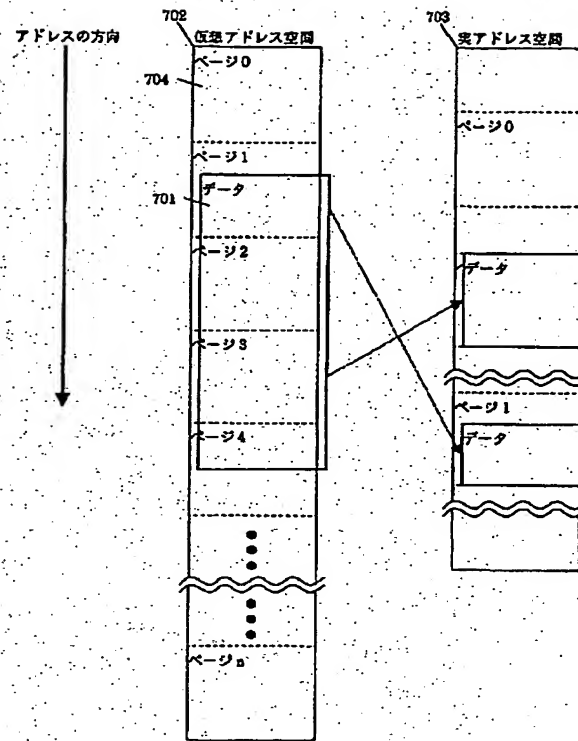
【図 8】

8



【図 9】

図 9



フロントページの続き

(51) Int. Cl.⁷
G 0 6 F 15/163

識別記号
6 5 0

F I
G 0 6 F 15/163

ターマコード (参考)
6 5 0 A